統計的機械学習 （応用計量分析２）第12回

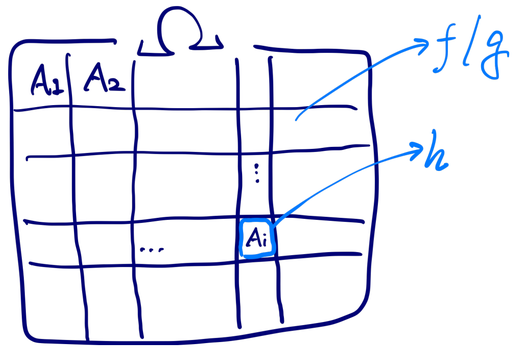
強化学習（参考pdf 17章）

補足：サヴェッジの公理のP3とP6の無矛盾性

出欠コード
XXXX

混ぜることで選考が変化する/しない公理で矛盾？→しない

- P3 単調性 結果 $y, y' \in \mathcal{Y}, y \succsim y' \quad \text{iff } f_A^y \succsim f_A^{y'}$
  - f のAにおける結果をy, y'にそれぞれ置き換えた行動の選好は、置き換えた結果の選好関係そのものになる
  - $f \sim f$であるので、Aをy, y'で置き換えたことによって$\sim$が$\succsim$に「変化した」ように思えるかもしれないが、$\succsim$は$\sim$を含むより弱い主張なので変化したわけではない
    - $3 = 3$だが、$3 \geq 3$も正しい
- P6 連続性 $\forall f, g, h \in F \text{ s.t. } f \succ g, \exists \{A_i\}_i^n : \text{split of } \Omega, f_{A_i}^h \succ g \text{ and } f \succ g_{A_i}^h$
  - (“確率”が)十分小さい事象に分割すれば、
    その1つに関して別の行動 h に置き換えても選好に影響しない
  - $|\Omega| = \infty$を含意、非原子的（どこまででも分割可能な）測度
  - A_iを置き換えることで$\succ$が$\prec$に変化しない**ようにできる**分割$\{A_i\}$が存在
- 構成的証明：主観的期待効用最大化に従う意思決定者は上記の両方を満たす
  - $f \succsim g$ iff $\int_\Omega u(f(\omega))\mathrm{d}\mu(\omega) \geq \int_\Omega u(g(\omega))\mathrm{d}\mu(\omega)$ となるようなuとμを持っている意思決定者



確率未知（曖昧性）の下での意思決定理論を扱った

- 不確実性には２種類ある
  - 客観的確率で表される不確実性：リスク（偶発的不確実性）
  - 確率自体の不確実性：曖昧性（認識論的不確実性）
- 確率が未知でも、それぞれ一定の公理のもとで以下がいえる
  - 主観的期待効用理論：公理P1—7に従う意思決定者は、主観的な確率と効用を持っていて、主観確率に関する期待効用を最大化している・すべき
  - マキシミン期待効用理論：公理AA1,2,3’,4,5および曖昧性回避の公理に従う意思決定者は、確率分布の凸集合を信念として持ち、その中の最悪ケースの期待効用を最大化している・すべき

本日の内容

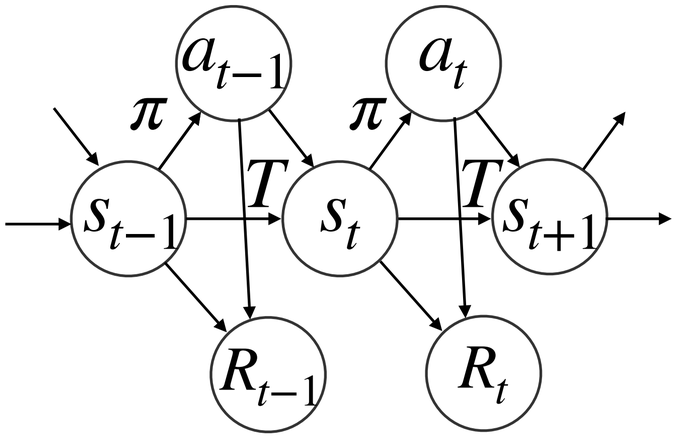
機械学習における逐次的意思決定、強化学習を扱う

- 1. ガイダンス・因果推論と機械学習の概論
- 2. 意思決定理論（期待効用理論）の復習、因果推論との関係
- 3. 潜在結果モデルに基づく因果推論の枠組み
- 4. 平均因果効果の推定法
- 5. 条件付き平均因果効果（CATE）の推定法１：メタ学習器
  - CATEの推定法２：二重機械学習
- 6. CATEの推定法３：決定木と決定森
  - 深層学習に基づく方法
- 7. 構造方程式モデルとバックドア基準
- 8. 因果探索
- 9. 発展的な因果推論手法：フロントドア調整、操作変数法
- 10. 発展的な因果推論手法：代理変数法、回帰不連続デザイン
- 11. 発展的な意思決定理論
- **12. 強化学習**
- 13. オフライン強化学習
- 14. バンディット

出欠コード
XXXX

マルコフ決定過程によるデータ生成を仮定

- **マルコフ決定過程**（Markov Decision Process; **MDP**）とは、以下の5つの要素から構成される意思決定問題
  - $\mathcal{S}$は環境のありうる状態の集合で、**観測される**
    - 観測されない状態のある問題を部分観測マルコフ決定過程というが、本講義では割愛
  - $\mathcal{A}$はとりうる行動の集合
  - 遷移確率 $T(s, s', a) = p(s'|s, a)$ ：状態$s \in \mathcal{S}$で行動$a \in \mathcal{A}$をとったとき次の時刻に状態s'に遷移する確率
    - 時刻tにおける状態確率が前時刻のみで書けるという**マルコフ性**を仮定： $p(s_t|a_{t-1}, s_{t-1}, \ldots, s_0) = p(s_t|a_{t-1}, s_{t-1})$
    - 1時刻差のみ扱うことが多いため、次の時刻の変数を x' のようにプライム（’）付きで表す
  - $R(s, a)$ は報酬関数：長期的にその総量を最大化すべき対象
    - おおむね効用に相当。確率変数 $R \sim p(R|s, a)$ でもよいが、簡単のためここでは決定論的とする
  - $\gamma \in [0,1]$ は割引率と呼ばれ，遠い将来の報酬を割り引いて考える度合い
- 以下の割引累積報酬を最大化する方策$\pi(a_t|s_t)$を得ることが目標
  - $J(\pi) = \mathbb{E}_{(s_t,a_t)_{t=0}^{T}}\left[\sum_{t=0}^{T} \gamma^t R(s_t, a_t)\right]$

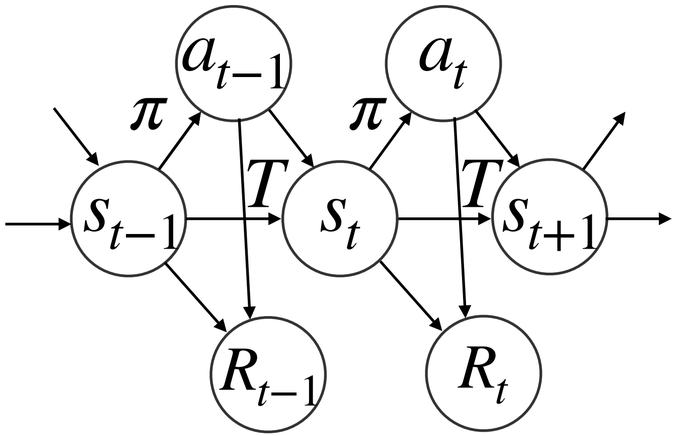


行動価値関数（Q関数）

出欠コード
XXXX

時刻tから最後までの評価値をその時刻の(s, a)の評価値として定義

- $J(\pi)$を最大化したいが、すべての時刻を同時に扱うのは大変　$J(\pi) = \mathbb{E}_{(s_t,a_t)_{t=0}^{T}}\left[\sum_{t=0}^{T}\gamma^t R(s_t, a_t)\right]$
  - πはその時刻の報酬を最大化すればいいわけではなく、**長期的に報酬が高まりそうな状態に遷移する**ことも目指す必要
    - 例）オセロで短期的に枚数を増やすよりも角を取る戦略
  - →途中の計算を「メモ化」したい（cf 動的計画法）
- 「ある時刻 t に行動 a をとり、そのあと π に従って T まで実行」した**行動価値**（部分割引報酬）を定義
  - $Q_t^{\pi}(s_t, a_t) := \mathbb{E}_{(s_{t'},a_{t'})_{t'=t+1}^{T}}\left[\sum_{t'=t}^{T}\gamma^{t'-t}R(s_{t'}, a_{t'}) \,\middle|\, s_t, a_t\right]$
- $\gamma < 1$が十分小さい、ないしTが十分大きければ$T = \infty$と近似できる
  →時刻tの依存性をなくせそう。$Q^{\pi}(s, a)$と書く
- Jとの関係： $J(\pi) = \mathbb{E}_{(a_0,s_0)}\left[Q^{\pi}(s_0, a_0)\right]$



Q関数の再帰性

ある時刻の$Q^\pi(s_t, a_t)$はその時刻の報酬と次の時刻の$Q^\pi(s_{t+1}, a_{t+1})$の割引期待値

- 行動価値は再帰的な構造を持つ

$$
\begin{aligned}
Q^\pi(s_t, a_t) &:= \mathbb{E}_{(s_{t'}, a_{t'})_{t'=t+1}^{\infty}} \left[\sum_{t'=t}^{\infty} \gamma^{t'-t} R(s_{t'}, a_{t'}) \;\middle|\; s_t, a_t \right] \\
&= R(s_t, a_t) + \gamma \mathbb{E}_{(s_{t'}, a_{t'})_{t'=t+1}^{\infty}} \left[\sum_{t'=t+1}^{\infty} \gamma^{t'-(t+1)} R(s_{t'}, a_{t'}) \;\middle|\; s_t, a_t \right] \\
&= R(s_t, a_t) + \gamma \mathbb{E}_{(s_{t+1}, a_{t+1})} \left[Q^\pi(s_{t+1}, a_{t+1}) \;\middle|\; s_t, a_t \right]
\end{aligned}
$$

- → 時刻tに状態sであり、行動aを行う「価値」$Q(s, a)$は、
  その時刻に得られる報酬 $R(s, a)$ と、それによる次の状態 $s' \sim p(s'|s, a)$ と次の行動 $\pi(a'|s')$
  に関して次の状態価値 $Q(s', a')$ の期待値のγ割引として陰に定義される
  - →この方程式を解けば Q^π がわかる
- しかし・・・

最適な行動の価値関数Q^*

最適な行動方策に関する価値関数Q*はQ*の最大化として再帰的に書ける

- しかし我々は行動方策πを固定してQを知りたかったのではなく、$J(\pi) = \mathbb{E}_{(a_0,s_0)}\left[Q^{\pi}(s_0, a_0)\right]$を最大化する方策$\pi$を知りたかったはず
  - 真の最適な方策 $\pi^* := \arg\max_{\pi} J(\pi)$ を考え、それについての行動価値$Q^{\pi^*} =: Q^*$を考える
- π^*は各時刻で最適な行動をとるので、決定論的関数として書ける
  - $\pi^*(s) = \arg\max_{a} Q^*(s, a)$
- これを前ページのQ^{π}の再帰式に代入すると
  - $Q^*(s, a) = R(s, a) + \gamma \mathbb{E}_{s'}\left[\max_{a'} Q^*(s', a') \;\middle|\; s, a\right]$
- Qとは無関係の未知のπを排除し、Qと観測可能な$R(s, a)$, s'のみから定まる式が得られた

Q学習の損失

データとの整合をとるTD誤差でQ*を学習

- 最適価値の定義から、両辺の不整合を損失に
  - $$Q^*(s,a) = R(s,a) + \gamma \mathbb{E}_{s'}\left[\max_{a'} Q^*(s',a') \;\middle|\; s,a\right]$$
- 時間差誤差（Temporal Difference；**TD**）
  - $$\delta := R(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a)$$
    - s'に関しては期待値の代わりにサンプリングされた値を用いている
- ある時刻の状態・行動と次の時刻の状態 s, a, s'（および$r = R(s,a)$）のみから定義される誤差が得られた

Q学習のモデルと学習法

テーブルベースと関数近似があり、関数近似の場合はmaxの部分は固定するFitted-Q

- 状態と行動の空間が有限離散的（$|\mathcal{S}|, |\mathcal{A}| < \infty$）なら、$Q(s,a)$を表として定義して
  - データから時間差ペア(s, a, s', r)を取得し、TD誤差δを計算し
  - $Q(s,a) \leftarrow Q(s,a) + \alpha_k \delta$　（α_kはkステップめの学習率）
  - のように順次誤差を修正すればよい
- $Q(s,a)$を関数としてNN等で学習する場合、TD誤差を現時刻の$Q(s,a)$に関してだけ微分をとる**Fitted-Q学習**を用いて学習できる

  $$y \leftarrow r + \gamma \max_{a'} Q_\theta(s', a')$$

  - $$\theta \leftarrow \theta - \alpha_k \frac{\partial}{\partial \theta} \left(y - Q_\theta(s,a)\right)^2$$
  - $\max_{a'}$を通して微分できないため
  - ※ 収束保証はQ_θが線形など限られた場合だけ。NNを含む一般には収束保証がないが、実用上はよく用いられる

（参考）強化学習のアプローチの分類

Q学習は方策オフでモデルフリーで価値ベースの手法

- 強化学習にはQ学習の他にもいろいろある
- 方策オン/オフ
  - 実際に取られた方策πを前提としたQ^{π}を学習するSARSA法は

    $\max_{a'}$の代わりに実際に取られたa'を用いる
- モデルフリー/モデルベース
  - Q学習は陽にT（や$R(s,a)$）をモデル化せず、

    「とにかく状態sのときaをするとどのくらい良いか」を推定する方法
  - Tを学習するモデルベースのほうがサンプル効率は高い傾向がある

    一方、近似誤差が累積するためTのモデルが真のクラスを含んでいないと問題
- 価値ベース/方策ベース
  - 直接的にπを最適化する方法もあり、特に行動が連続的なロボットなど$\arg\max_a Q(s,a)$がリアルタイムで

    実行しづらい場合によく用いられる（ただしサンプル効率はより悪い傾向）
  - 方策と価値を同時に学習する方法も：Actor-Critic

出欠コード
XXXX

強化学習の典型的な定式化であるMDPとその学習法Q学習

- 強化学習の定式化MDPは状態、行動、遷移確率、報酬、割引率から定義される
- 各時刻の状態と行動に対する価値の評価値であるQ関数を定義
  - Q関数は再帰性があり、次の時刻の報酬とQ関数だけを用いて（2ステップ先以降を用いずに）書ける
- 再帰性の1時間ステップ差の整合性（TD誤差）から損失関数を定義
- TD誤差最小化としてQ学習を構成